**C S 487/519 Applied Machine Learning I**

# Compare regression methods

1. **Objective**

In this *individual* homework, you are required to understand and compare several regression algorithms.

# Requirements

* + (60 points) Write code to conduct regression by
    - (30 points) utilizing several regression functions: (i) LinearRegression, (ii) RANSACRegressor, (iii) Ridge, and (iv) Lasso. These functions are provided by the Python scikit-learn library.
    - (10 points) implementing the normal equation solution, and
    - (20 points) using one approach to conduct non-linear regression.

(20 points) Each regressor (except the normal equation method) needs to be tested using two datasets:

*•*

(1) the *housing* dataset, which can be downloaded from [here](https://www.cs.nmsu.edu/~hcao/teaching/cs487519/data/housing.data.txt), and (2) the California Renewable Production 2010-2018 dataset which can be downloaded from [here](https://www.kaggle.com/cheedcheed/california-wind-power-generation-forecasting/data). For the Normal equation solution, you just need to test it using the housing dataset.

(15 points) Properly analyze the regressors’ behavior by applying the knowledge that we discussed in class. Such analysis should include at least Mean squared error (MSE) (or R2 score, or residual plots) and running time. Put the analysis to **report.pdf** file.

*•*

* + (5 points) Write a readme file **readme.txt** with the commands to run your code. Your code needs to run in command line, accepting as input parameters the algorithm name, the dataset filename, and any required parameter. For example, “python main.py ridge dataset.csv –alpha 1.0”

(10 points; **Bonus**) Design a new approach to improve the existing regressor. You need to describe your method in detail in the report file. Your analysis should also include this new method. If your new approach outperforms the existing methods in a reasonable way (e.g., largely decrease MSE while using similar amount of running time, or generate similar MSE but reduce the running time dramatically), you will get the bonus points. If your new method does not outperform the existing method, you will only get two bonus points.

*•*

* + Your Python code should be written for Python version 3.5.2 or higher. Please properly organize your Python code (e.g., create proper classes, modules). Each required task had better be implemented in a separate python file and imported into the main script. For example, to use the Lasso regression method you can create the script **mylasso.py**, then use “import mylasso” in the main.py file to test your implementation.

# Submission instructions

Put all your files (Python code, readme file, report, etc.) to a zip file named **hw.zip** and upload it to Canvas.

# Grading criteria

* + 1. The score allocation has already been put beside the questions.
    2. Please make sure that you test your code **thoroughly** by considering all possible test cases. Your code may be tested using more datasets. **For this homework, your code will not be tested using more datasets. Thus, it does not need to be flexible to accept different datasets as input**.
    3. 5 points will be deducted if submitted files (including files types, file names, etc.) do not follow the instructions.
    4. If the total points are more than 100. Your grades will be scaled to the range of [0,100].